Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 1275 – 1284

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2016, 5-7 September 2016, York, United Kingdom

Visualization of Superficial Similarities between Data Jackets for Aiding Creativity on Innovators Marketsplace on Data Jackets

Norisada Masui[a]*, Yukio Ohsawa[b]

[a]*Department of Systems Innovation, School of Engineering, The University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, 113-8656*
[b]*Department of Systems Innovation, School of Engineering, The University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, 113-8656*

Abstract

In recent years, various data has been obtained at different situations due to the improvement of the sensors or the popularization of the Internet and mobile phones. In the field of business, academia or politics, the demand for supporting decision making based on data has been increased. Ohsawa et al. propose the concept of the market of data (MoDAT) for the demand, and provide Innovators Marketplace on Data Jackets (IMDJ) as a method for realizing MoDAT, which is a workshop method for encouraging particippants to utilize and exchange data. However, although participants of the workshop discuss the given theme each other, possible combinations of DJs for satisfying requirements are hardly generated. That is one of the problems of IMDJ. In this paper, in order to solve the problem, we propose the visualization method for representing the similarities between DJs using latent dirichlet allocation (LDA) and multidimensional scaling (MDS) in order to aid users metacognition in IMDJ. The result shows that our proposed method improve the performance of participants to create feasible solutions by combining DJs.

Peer-review under responsibility of KES International

Keywords: Information visualization, Metacognition, Creativity, Innovators Marketplace on Data Jackets, Latent Dirichlet Allocation, Multidimensional Scaling

1. Introduction

A few years have passed since the term ”big data” come to be recognized. The technical report of McKinsey Global Institute pointed out the seven features of big data in 2011[1]. These are almost observed in the actual world nowadays.

- Data have swept into every industry and business function and are now an important factor of production.
- Big data creates value in several ways.

* Corresponding author. Tel.: +81-3-5841-2908 ; fax: +81-3-5841-2908.
E-mail address: 1083505752@mail.ecc.u-tokyo.ac.jp

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.172

- Use of big data will become a key basis of competition and growth for individual firms.
- The use of big data will underpin new waves of productivity growth and consumer surplus.
- While the use of big data will matter across sectors, some sectors are poised for greater gains.
- There will be a shortage of talent necessary for organizations to take advantage of big data.
- Several issues will have to be addressed to capture the full potential of big data.

At the same time, governments of countries adopted the policy called Open Government, which allow citizens to access governmental records. Under this situation, the movement called Open Data is also getting popular. The movement of encourage governments to Open Data is the idea that data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. The web services, e.g., data.gov and data.jp, have been published online.

In addition to Open Data, public and private sectors expect to take advantage of data stored in various domains. However, although they recognize the importance of data utilization and exchange, the opportunities to share their data, to collaborate each other and to create new business have been limited in Japan. Under these circumstances, Ministry of Economy, Trade and Industry (METI) established the conference for data-driven innovation in 2014, in order to lead innovation via data utilization and exchange, The organizations such as Data Exchange Consortium (DXC) are established in order to create new businesses by exchanging data among companies. These facts may indicate the increase of expectations for making decisions by utilizing data in various fields.

Although the expectation for taking advantage of data have been increasing, there are some barriers to achieve it. For example, the media reported that East Japan Railway Company seriously violated the privacy of passengers by selling the commuting history to other companies and these companies were severely criticized by public opinions. The problem gave us the chance to consider about how we can create the framework to promote decision making based on data without the fear of privacy issue. Nowadays, although there are services which provide the platform to exchange data such as Microsoft Azure Marketplace, CKAN or KDnuggets, they can not serve a function as the platform where data owners and users communicate each other and make innovations because they only exhibit superficial information of data without communication among stakeholders. In such a situation, Ohsawa et al. proposed the market of data (MoDAT)[2]. The concept of MoDAT is different from these services. It is designed as the place where each owner and user of data can learn the value of each database/dataset, and each of them can buy and sell database/dataset in a reasonable condition.

Innovators Marketplace on Data Jackets (IMDJ)[3,4] is an approach for realizing MoDAT. IMDJ is a gamified workshop for leading innovative collaboration of data. Stakeholders include data owners, users and analysts, and they communicate to create the solutions to satisfy the users requirements by combining DJ[5], a key concept of this method. DJ is a summary of datasets including various information, e.g., the title, the explanation, the variable labels, the owners, information about sharing policy and data format. Currently 914 DJs have been registered (April, 2016). This technique is essential for MoDAT because data owners do not have to open their own data, but the information about their data, which reduce the fear of loss of business opportunities and violation of privacies. The process of IMDJ consists of following four steps. 1) Data owners register their dataset as DJs. 2) Possible combinations of DJs are visualized as a scenario map for supporting participants to discover latent combinations of each dataset using some visualization tools, e.g., KeyGraph[6]. 3) Participants create ideas by combining DJs for satisfying data users' requirements. 4) Participants evaluate their datasets through evaluating the ideas created by combining DJs. The process of IMDJ encourage participants to negotiate for exchanging or analyzing data, which activates MoDAT.

IMDJ has been introduced in various themes, and succeeded in supporting participants to exchange their data and take actions. However, some problems are pointed out. One of the problems is that although a lot of ideas for satisfying requirements are created in IMDJ, they have been hardly realized as real businesses. The aim of this study is to improve this situation by proposing new method to visualize the relationships of DJs.

2. Related Work

In this chapter, we first introduce KeyGraph as the existing visualization method most frequently employed in IMDJ and then introduce the studies in information retrieval and information visualization, which are related fields to our proposed method. KeyGraph[6] has been originally proposed as an algorithm for extracting keywords from documents.

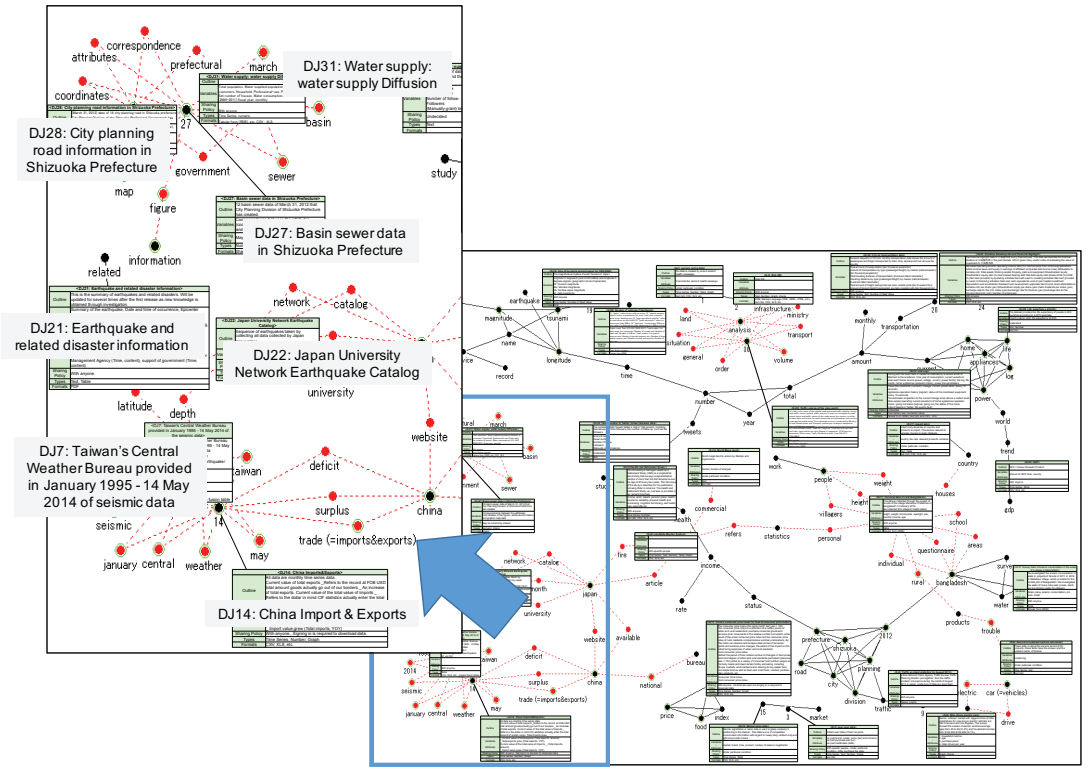

Fig. 1. Visualization of relationships between DJs by KeyGraph.

Ohsawa et al. proposed KeyGraph and applied it to various areas of study such as marketing or earthquake prediction based on the concept of chance discovery [7]. In IMDJ, each DJ is represented as a black node and each key factor that bridge between DJs, extracted from the texts of DJs, are represented as a red node. Fig. 1 shows one example of visualized map on which related DJs are linked via the keywords co-occurred in DJs, using KeyGraph.

In this paper, we deal with the texts, names of the variables or words contained in explanations, in DJs. So it is important to introduce the study of information retrieval (IR) related to texts. The importance of IR has been recognized since about 1000years ago. In a few decades, the development of computers cause to improve the quality of studies in this fields [8]. The vector consisting of the frequencies of terms, called bag-of-words, that occur in a document has been widely used in order to represent the features of a document. This representation of documents has been applied to a lot of models of language. Tf-idf, a prevalent method of IR, is based on the representation of bag-of-words, as defined below.

$$(tf \cdot idf)_{i,j} = tf_{i,j} \cdot idf_i$$

$$\left(tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \right) \tag{1}$$

, where $n_{i,j}$ is the number of times that term i occurs in document j, $|D|$ is total number of documents in the corpus, and $|\{d : t_i \in d\}|$ is the number of documents in D where term i appears. Tf-idf scheme works in extracting the features of documents, but not efficient in reducing dimensionality of features. Topic model is statistical language model, a probability distribution over sequences of words, assuming the existence of latent variables, which is efficient for both extracting the features of documents and efficient on reducing dimensionality of features. It includes unigram model, n-gram model, latent semantic indexing (LSI), probabilistic LSI or latent Dirichlet allocation (LDA). In this paper,

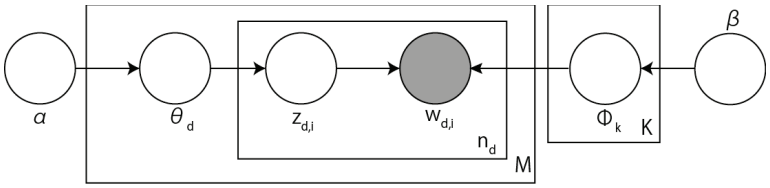

Fig. 2. Graphical model of LDA.

we use LDA for our method. In addition to the merit of topic model itself, LDA is superior to other models in which topic is hypothesized in the view of the amount of calculation and versatile applicability[9].

LDA[10] is a Bayesian probabilistic model of text documents. Fig. 2 shows the graphical model representing the generative process of LDA. Let us think the corpus **D** including M documents. First, we define document d as a sequence of n_d words denoted by $\mathbf{w}_d = (w_{d,1}, w_{d,2}, ..., w_{d,n_d})$, where $w_{d,i}$ is the ith term in the sequence. So the corpus is represented as $\mathbf{D} = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M)$. LDA assumes each document has some topics, latent variables for characterizing documents. That is, the ith term of document d is assumed to generated by topic $z_{d,i}$. When $\theta_{d,k}$ is the probability that topic k appears in document d, we define the topic distribution of document d as $\boldsymbol{\theta}_d = (\theta_{d,1}, \theta_{d,2}, ..., \theta_{d,K})$, where K is the constant representing the number of topics. In addition, when $\phi_{k,v}$ is defined as the probability that term v appears in topic k, we define the word distribution of topic k as $\boldsymbol{\phi}_k = (\phi_{k,1}, \phi_{k,2}, ...\phi_{k,V})$, where V is the number of unique terms. $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ is assumed to follow the generative process by a Dirichlet distribution.

$$\begin{aligned} \boldsymbol{\theta}_d &\sim Dirichlet(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \\ \boldsymbol{\phi}_k &\sim Dirichlet(\boldsymbol{\phi}_k|\boldsymbol{\beta}) \end{aligned} \tag{2}$$

, where the parameter $\boldsymbol{\alpha}$ is a K-vector denoted by $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)(\alpha_k > 0)$ and the parameter $\boldsymbol{\beta}$ is a V-vector denoted by $\boldsymbol{\beta} = (\beta_1, ..., \beta_V)(\beta_v > 0)$. And $w_{d,i}$ and $z_{d,i}$ are assumed to be generated by a multinomial distribution.

$$\begin{aligned} z_{d,i} &\sim Multinomial\,(z_{d,i}|\boldsymbol{\theta}_d) \\ w_{d,i} &\sim Multinomial\left(w_{d,i}|\boldsymbol{\phi}_{z_{d,i}}\right) \end{aligned} \tag{3}$$

A Dirichlet distribution and a multinomial distribution is defined as below. First, a k-dimentional Dirichlet random variable $\boldsymbol{\theta}$ ($\theta_i \geq 0, \sum_i^k \theta_i = 1$) has the following probability density, a Dirichlet distribution.

$$Dirichlet(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k \theta_i^{\alpha_i - 1} \tag{4}$$

, where $\boldsymbol{\alpha}$ is a k-vector with components $\alpha_i \geq 0$ and $\Gamma(x)$ is Gamma function. As for a multinomial distribution, we may think n independent trials, for example, casting dices, each of which leads to a success for one of K categories, where each category having a given fixed success probability denoted by $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_K)(\sum_{k=1}^K \phi_k = 1, 0 \leq \phi_k \leq 1)$. A multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories.

$$Multinomial\left(\{n_k\}_{k=1}^K|\boldsymbol{\pi}, n\right) = \frac{n!}{\prod_{k=1}^K n_k!} \prod_{k=1}^K \pi_k^{n_k} \tag{5}$$

, where n_k represents the number of appearance of category k. If LDA is used in practice, the parameters should be inferred with the corpus given. There are some methods to infer the value of parameters of LDA from corpus, e.g. Gibbs sampling or collapse Gibbs sampling as sampling approach or variational Bayes (VB) or collapse VB as Bayesian approach. In the paper, we adopt online VB[11] as an inference method.

Although KeyGraph is mainly used in IMDJ focusing on the text data, a lot of methods of visualization have been proposed in domains. Information visualization (InfoVis) is a research area related to KeyGraph. InfoVis aims to aid users in exploring, understanding, and analyzing data through progressive, iterative visual exploration.

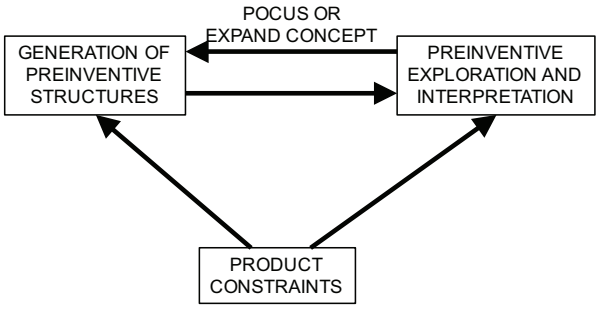

Fig. 3. Geneplore Model[13].

Liu et al.[12] classify the studies of InfoVis into four categories (empirical methodologies, interactions, frameworks and applications). KeyGraph and our proposed method may be classified the category of applications. Researcher engaged in this category try to apply various InfoVis techniques to different fields. Here, we discuss text visualization, a type of applications, which is most related to our proposed method introduced in the next chapter. There are several text visualization application using LDA.

3. Proposed Approach for Visualization to Aid Metacognition

As mentioned above, the problem is pointed out that although a lot of ideas for satisfying requirements are created in IMDJ, feasible combination of DJs are hardly generated. This paper aim to solve this problem by proposing a new method to visualize the relationship of DJs. In this chapter, we discuss the concept of the proposed method, particularly from a point of view of cognitive science and then we explain the detail of the proposed method.

3.1. Concept

IMDJ is the process of problem solving. Let us think of the situation of IMDJ where you get some users requirements. We should create solutions to satisfy these requirements in the workshop. The requirements we are trying to satisfy seldom have specified goals because the requirements may be vague or unclear. Therefore we have a lot of solutions to satisfy the requirements, although the qualities of solutions may be various. In order to support this complex process, we introduce the theory of creative cognition, metacognition and analogy.

Finke et al. propose the Geneplore model[13] as the fundamental of creative thinking. This model assumes the cycle of two subprocesses, generative process and exploratory process for creative cognition. In the first phase, we construct mental representation called preinventive structures, having various properties that promote creative discoveries. The preinventive structures can be thought as internal precursor of the final, externalized creative products. In the second phase, we seek to interpret the preinventive structures in meaningful ways, exploiting the properties the preinventive structures have. If we do not satisfy the preinventive structures, we can go back to the initial phase. The preinventive structures would be generated, regenerated, and modified through the cycle until the conceptual refinement or extension of the preinventive structures achieve desired level. In addition, constraints on the creative products are considered to affect the underlying cognitive process in the cycle. Fig. 3 shows the structure of this model as is mentioned above and Table 1 shows examples of the key factors of the Geneplore Model. They also say that there are some creative strategies for problem solving. That is important for thinking of IMDJ because IMDJ is the process of problem finding and solving, in which creativity is essential. Here, We discuss metacognition and analogical reasoning.

Metacognition is defined as the mind's ability to monitor and control itself or, in other words, our ability to know our knowing. Nelson and Narens construct classical metacognitive model, where metacognition is considered as the interplay between two levels of information processing, a single meta-level and a single object-level[14]. In this model, it is assumed that the object-level monitors the meta-level, evaluating the information from object-level, and the meta-level control the object-level, revising the object-level activities based on the evaluation. It is widely accepted that

Table 1. Example of Cognitive Processes, Structures, Properties, and Constraints in the Geneplore Model [13].

Generative Processes	Preinventive Structures	Preinventive Properties	Exploratory Processes	Product Constraints
Retrieval	Visual patterns	Novelty	Attribute finding	Product type
Association	Object forms	Ambiguity	Conceptual interpretation	Category
Synthesis	Mental blends	Meaningfulness	Functional inference	Features
Transformation	Category exemplars	Emergence	Contextual shifting	Functions
Analogical transfer	Mental models	Incongruity	Hypothesis testing	Components
Categorical reduction	Verbal combinations	Divergence	Searching for limitations	Resources

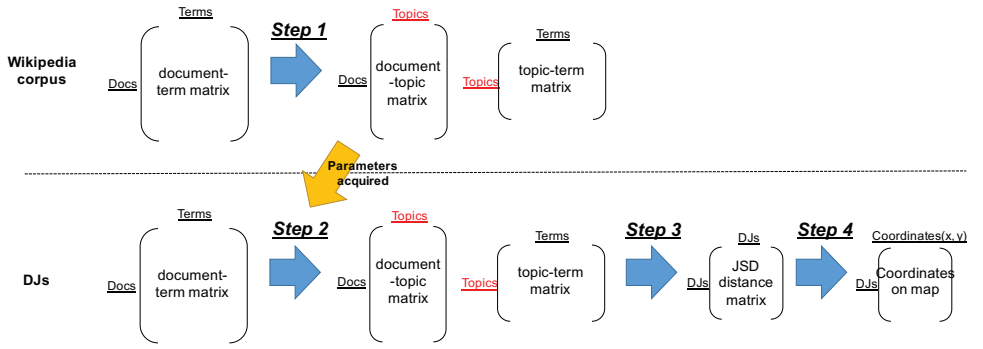

Fig. 4. Steps of proposed method.

metacognition is effective on creative problems solving. For example, Ormond et al. indicated that a youth who have rich metacognitive knowledge for decision making is likely to make superior decisions [15].

Analogical reasoning is a cognitive process of transferring information or meaning from a particular subject (the source) to another (the target). The theory of structure mapping proposed by Gentner [16] and the theory of analogical mapping by constraint satisfaction proposed by Holyoak [17] are most established as the theory of analogy. Analogy is broadly considered essential for problem solving.

What kind of visualization is proper to achieve the goal to solve the problem that feasible combination of DJs are hardly generated in IMDJ? Our policy for creating new method based on three cognitive abilities mentioned above is that if visualization created by the proposed method is used on IMDJ, 1) participants can recognize their own specialization or their expert knowledge and discuss the ideas to satisfy users' requirements in IMDJ as taking them into considerations, and 2) apply ways of problem solving in other fields to creating ideas, i.e. analogical thinking from other's opinions.

3.2. Approach of our method

Our proposed approach for visualization consists of following four steps. Fig. 4 shows the steps of proposed method easily. We explain each step detailedly.

1) The first step is to make the topic space using external information. We used the Japanese wikipedia corpus consisting of 761,989 documents including 16,197 unique terms (April, 2016) as external information. In addition, we also used the corpus using all DJs in order to compare the result of the experiments explained in the next chapter. First, we make an document-term matrix by doing the morphological analysis by MeCab, a tool for morphological analysis of Japanese documents, selecting terms according to the parts of speech (a kind of verb and adjective) and counting the frequency of each term in each document. Then we calculated the document-topic matrix and the word-topic matrix by LDA. The number of topic is experimentally decided to be 50 in the model of wikipedia and 70 in the model of DJs for comparison.

2) The second step is to map DJ space into topic space. In other words, we infer the topic distributions of each DJ. We make the text representation on each DJ. Then this collection of texts are changed into an document-term matrix as the same way of the step 1. Finally, each DJ is mapped into the Wikipedia space. In other words, using

the parameters acquired in the step 1, we make the topic distribution of each DJ. The algorithm to infer is online VB written in Section 2. For implementing this algorithm, we use the python package, gensim.

3) The third step is to choose the DJs to be used in the workshop and calculate the distances between DJs from their topic distributions. First, we chose about 30 DJs related to the theme of the workshop and calculate the distance matrix, containing the distances between DJs calculated from the topic distributions of selected DJs. In this paper, Jensen Shannon divergence (JSD) , which indicates the closeness of two distributions, is adopted as a measure of distance. We select N DJs for the workshop and use the matrix obtained in the second step, where the number of topics is K. The topic distribution of $DJ_n (1 \leq n \leq N)$ is represented as $\mathbf{P}_n = (p_{n,1}, ..., p_{n,K})$, where $p_{n,k}$ is the probability that DJ_n belongs to the topic k ($0 \leq p_{n,k} \leq 1$). JSD between DJ_i and DJ_j ($1 \leq i, j \leq N$) is calculated bellow.

$$JSD_{i,j} = \frac{1}{2}\left(\sum_{k=1}^{K} P_{i,k} \log \frac{P_{i,k}}{\frac{P_{i,k}+P_{j,k}}{2}} + \sum_{k=1}^{K} P_{j,k} \log \frac{P_{j,k}}{\frac{P_{i,k}+P_{j,k}}{2}}\right) \tag{6}$$

4) The final step is to visualize the relationships of DJs on a two-dimensional map. In this process, it is necessary to transform the distance matrix representing JSD into the coordinates of each DJ on two dimensional map. In order to transform as saving the relationship of distances of each DJ on the map visualized, we adopt multidimentional scaling (MDS) as a method to reduce dimensionality. In MDS, mapping given distance matrix into m-dimentional MDS space (here, m is 2) is calculated by the mapping f, i.e., $f : JSD_{i,j} \mapsto d_{i,j}$, where $d_{i,j}$ is the Euclid distance between DJ_i and DJ_j on MDS space, the map visualized. In the paper, identity mapping is used as the mapping f. MDS seek the coordinate of each DJs in order to reduce the sum of squared errors between $f(JSD_{i,j})$ and $d_{i,j}$, called Stress. In order to minimize Stress, we use SMACOF algorithm, which consists of four steps: 1) the coordinates in MDS space is randomly decided and the coefficient ϵ as threshold is set, 2) Stress is calculated, 3) the coordinates are updated by Guttman Transform, 4) the step 2 and the step 3 are repeated and if the value of Stress become less than ϵ, the update process is finished. The coordinates of DJs is used to visualize the map. Fig. 5 shows the example of the visualization created by our method.

4. Experimental Evaluation

In this chapter, we conduct the experiment in order to evaluate proposed method. We prepared two themes: C) "To think of our future by thinking the development of cars" and S) "To think of the plan to improve the local life by agriculture", and used two types of corpus as is mentioned above: W) Wikipedia and D) the collection of all DJs. Therefore we made four types of map for the experiment by combining the theme (C or S) and the type of corpus (W or D). The number of subjects was 14, who all were in twenties and were consisted of 12 men and 2 women. In the experiment, each subject created three ideas for requirements given in advance as using the visualization, which was like IMDJ conducted by one person. Each subject did the experiment twice with different map. The data we acquired in this experiments is bellow.

- The combination of two DJs of which each subject felt the distance on the map strange
- Three Ideas
  - the concepts
  - the requirements to satisfy
  - the IDs of used DJs
  - the data to be added for realizing the idea
- Self-evaluation of each ideas (rank from 1 to 5, representing from strongly agree to strongly disagree)
  - I am confident in this idea.
  - This idea has originality.
  - I combine DJs well in order to make this idea.
  - This idea is based on some DJs shown on the map or added by myself.
  - I make good use of the map of DJs.

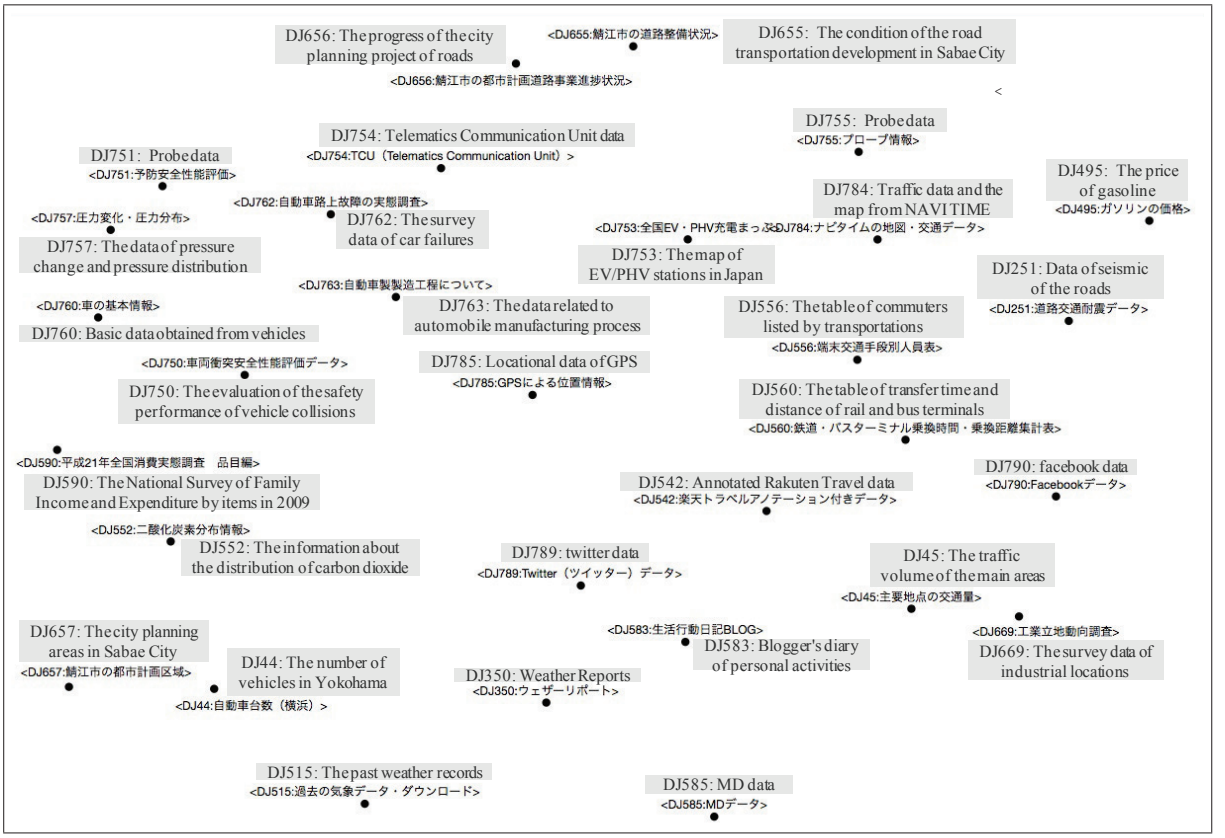

Fig. 5. One example of visualization by our method (The theme of workshop where visualization are used is ”To think of our future by thinking the development of cars”. If you want to access to more information about DJs used in Fig. 4 or information about other DJs, please visit http://www.panda.sys.t.u-tokyo.ac.jp/hayashi/djs/djs4ddi/.).

  - This idea is possible to realize by someone in the future.
  - This idea is related to my specialization.
  - This idea satisfy the given requirements.
- ID of DJ in which each subject specialize or is strongly interested, i.e. expertise

For example, let us see the ideas made by subjects participating in the experiment using the visualization whose theme is ”C” and corpus is ”W”. Here, in order to satisfy one of the given requirements, problems of the environmental pollution should be resolved, the idea that the price of gasoline for users of cars having less environmental impact should be decreased based on the degrees of the impact was suggested. This idea was based on DJ 495 (The price of gasoline), DJ 515 (The past weather records), DJ 552 (The information about the distribution of carbon dioxide) and DJ 760 (Basic data obtained from vehicles). Those who suggested the idea had specialized or been strongly interested in DJ 515 and DJ 760.

As we analyzed the created ideas and the acquired data, we focus on the two points of the results: 1) the ability of metacognition is important to create more feasible ideas, 2) proposed method work to show the similarities represented as distance. We explain the details respectively.

One of the results is that the participants who have better ability of metacognition may create more feasible ideas, which is consistent with the results of previous studies on metacognition. Here, the ability of metacognition may be considered as the ability to be conscious of their expertise and to use them in order to create ideas. That is because the self-evaluation of the idea is likely to be increasing if the idea is created with DJs which the subject specialize in. It

is proved by correlations, Table 2 shows, between the scores of self-evaluation and the number of DJs which subject specialize in. So we conclude that the ability of metacognition is important to create more feasible ideas.

Table 2. Correlation between points of self-evaluation and the number of DJs in which each subject specialize in.

Questionnaire Item	Correlation Coefficient
I am confident in this idea.	-0.315722
This idea has originality.	-0.104207
I combine DJs well in order to make this idea.	0.008801
This idea is based on some DJs shown on the map or added by myself.	-0.12419
I make good use of the map of DJs.	-0.124558
This idea is possible to realize by someone in the future.	-0.233661
This idea is related to my specialization.	-0.379582
This idea satisfy the given requirements.	-0.227052

Another is that proposed method work to show the similarities represented as distance. That is because the number of the strange combination of two DJs is less in the experiment using the visualization created by Wikipedia corpus than that in the experiment using the visualization created by DJ corpus. Here, the strange combination of two DJs means the combination of two DJs of which each subject consider the distance on the map strange. The two-way analysis of variance (ANOVA) verified the effects of both themes of workshops and corpus used in the process of the visualization for the number of the strange combination of two DJs ($p < 0.05$).

5. Conclusion and Future Scope

In this paper, the problem we try to solve is that although a lot of ideas for satisfying requirements are created in IMDJ, feasible combination of DJs are hardly generated. In order to solve the problem on IMDJ, We propose the visualization method for representing the similarities among DJs based on LDA and MDS. From the results of the experiments, we insist on two points. 1) The experiment for evaluation found that the participants who have better ability of metacognition may create more feasible ideas, consistent with the results of previous studies on metacognition. This consistent with the result of Finke's experiment that proper constraints on the creative products improve creativity. 2) As we expected, the proposed method work to show the similarities represented as closeness in the visualized map, which humans recognize closeness of the concepts of DJs intuitively. In other words, in our proposed method, DJs whose meanings are similar are located closely on the map of visualization. This supports the participants in IMDJ, because close located DJs are assumed to be categorized in the same fields, and the participants can grasp the location of the fields on the map where they specialized or are interested. The result suggests that our proposed method support participants of IMDJ to create the ideas reflecting the specialization of each participant and to execute analogical reasoning (particularly, the process of searching the base domain) for creating ideas to satisfying the requirements. In conclusion, the proposed method works better performance in solving the problem.

In the future work, we consider to introduce new rule that participants should always take their specialization into account in workshop we conduct. This may cause the result of workshop to become more effective, concrete and ingenious because the process of metacognition is promoted. In addition, it is necessary to make experiments in IMDJ introducing several participants, because the purpose of IMDJ is to encourage data utilization and exchange among various participants, which is achieved through the combination and analogical thinking caused by sharing expertise.

Acknowledgements

This study has been supported by JST CREST. I greatly appreciate it. And I am also deeply grateful to Mr. Teruaki Hayashi, my colleague of the laboratory, for great support to this research.

References

1. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big Data: The next Frontier for Innovation,Competition, and Productivity. McKinsey Global Institute; 2011.
2. Hayashi T, Ohsawa Y. Knowledge structuring and reuse system design using RDF for creating a market of data. Signal Processing and Integrated Networks (SPIN), 2015 2nd International Conference on. IEEE. 2015.
3. Ohsawa Y, Kido H, Hayashi T, Liu C, Komoda K. Innovators Marketplace on Data Jackets, for Valuating, Sharing, and Synthesizing Data. In: Tweedale JW, Jain LC, Watada J, Howlett RJ, editors. *Knowledge-Based Information Systems in Practice*. Springer International Publishing; 2015. p. 83-97.
4. Ohsawa Y, Liu C. Suda Y, Kido H. Innovators marketplace on data jackets for externalizing the value of data via stakeholders requirement communication. 2014 AAAI Spring Symposium Series. 2014.
5. Ohsawa Y, Kido H, Hayashi T, Liu C. Data jackets for synthesizing values in the market of data. *Procedia Computer Science* 2013;**22**:709-716.
6. Ohsawa Y, Benson NE, Yachida M. KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on. IEEE, 1998.
7. Ohsawa Y, McBurney P. *Chance Discovery*. Springer Berlin Heidelberg; 2003.
8. Singhal A. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin* 2001;**24.4**:35-43.
9. Nenkova A, McKeown K. A Survey of Text Summarization Techniques. In: Aggarwal CC, Zhai CX, editors. *Mining Text Data*. Springer Science and Business Media; 2012. 43-76.
10. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. *Journal of MachineLearning Research* 2003;**3**:993-1022.
11. Hoffman MD, Blei DM, Bach F. Online Learning for Latent Dirichlet Allocation. advances in neural information processing systems. 2010.(NIPS 2010) 856-864.
12. Liu S, Cui W, Wu Y, Liu M. A survey on information visualization: recent advances and challenges. *The Visual Computer* 2014;**30.12**:1373-1393.
13. Finke RA, Ward TB, Smith SM. *Creative cognition: Theory, research, and applications*. A Bradford Book; 1996.
14. Van Overschelde JP. Metacognition: Knowing About Knowing. *Handbook of Metamemory* 2008;**47**:47-71.
15. Ormond C, Luszcz MA, Mann L, Beswick G. A metacognitive analysis of decision making in adolescene. *Journal of Adolescence* 1991; **14**:275-291.
16. Gentner D. Structure-mapping: Theoretical framework for analogy. *Cognitive Science* 1983; **7.2**:155-170.
17. Holyoak KJ, Thagard P. Analogical mapping by constraint satisfaction. *Cognitive Science* 1989;**13**:295-355.
18. Borg I, Groenen PJF. *Modern multidimensional scaling: Theory and applications*. Springer Science+Business Media, Inc.; 2005.